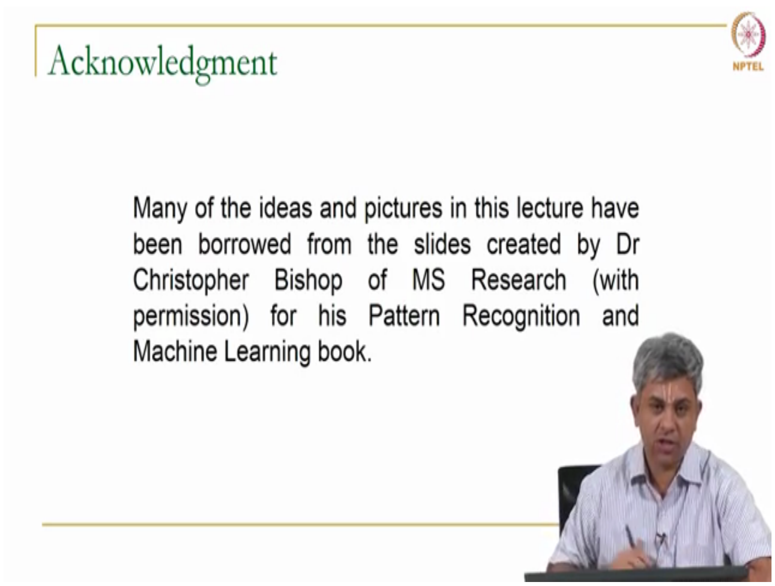
Machine Learning for Engineering and Science Applications
Professor Dr. Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
Conditional, Joint, Marginal Probabilities Sum Rule and Product Rule Bayes' Theorem

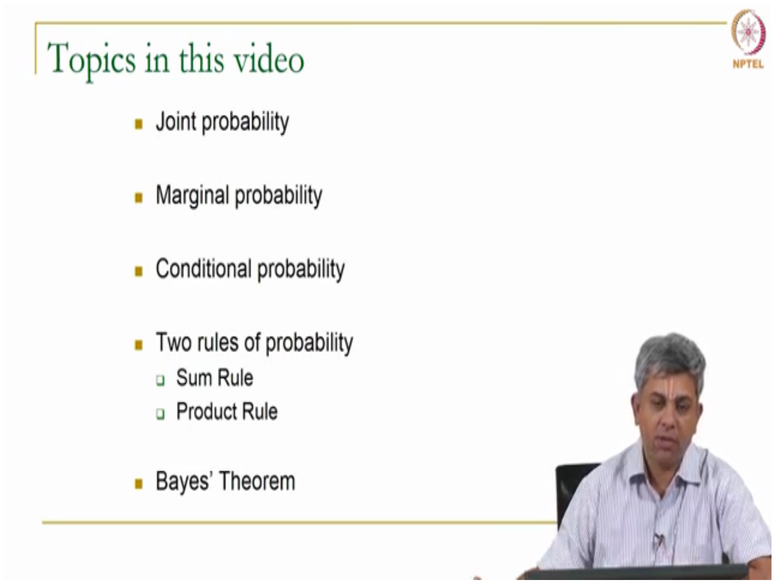
In this video we will be continuing with our discussion of probability theory. We will be looking at a few basic ideas beyond what we looked at in the last video. So we look at conditional, joint and marginal probabilities and two rules which essentially govern all of probability theory, the sum rule and the product rule. And finally we look at just the definition and a simple derivation of the Bayes' Theorem. We will look at Bayes' Theorem in greater detail and then next video after this one.

(Refer Slide Time: 0:48)



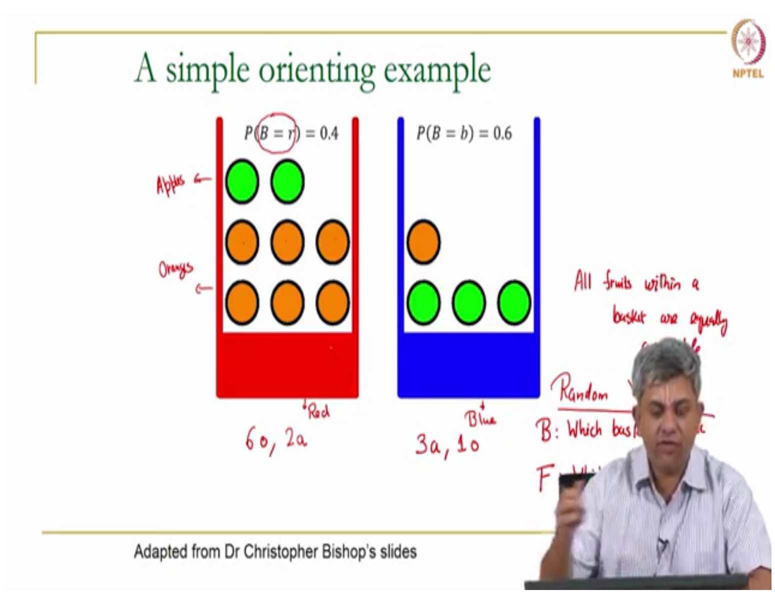
So today we are going to look at just the basics. So a quick acknowledgment, several of the ideas and the pictures in this lecture have been borrowed from the book by Dr. Christopher Bishop. You might remember that this one of the references for this course. The book itself is available freely on the web, legally freely available on the web. This has been made available by the courtesy of Microsoft Research, Dr. Christopher Bishop and his finger. I also want to mention that these slides and the pictures in the slides not the slides themselves have been, many of them have been borrowed from Dr. Christopher Bishop with his kind permission.

(Refer Slide Time: 1:27)



So the topics that we are going to look at, as I said before, are joint probability, marginal probability, conditional probability. All these three are simple ideas when you have more than one variable. We were looking at cases with one variable in the last video, now we are going to look at more than one random variable. And two of the rules that govern all of probability theory and finally Bayes' Theorem.

(Refer Slide Time: 1:48)

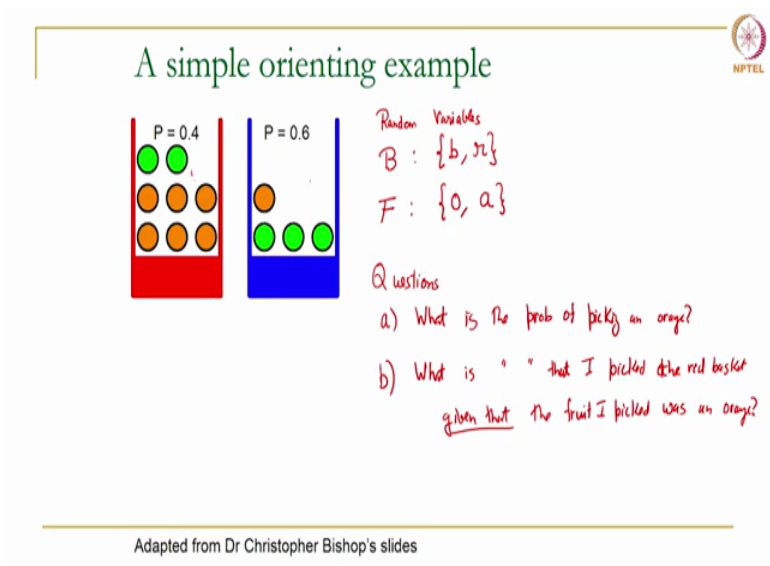


So here is a simple example. We will be using this example throughout this particular video. This is again from Christopher Bishop's book. So imagine that you have two baskets. One of these is a red basket and one of them is a blue basket. And each of these baskets has some fruits. The orange ones you can assume are oranges and the green ones we will assume are apples. Why green? Because the basket is red just for clarity. So let us say we have these two baskets and our task is to randomly put our hand into one of the baskets and pick out a fruit.

So let us say that all fruits within a basket are equally available. Even though for clarity we have drawn as one fruit is at the top and few others are to the bottom. Assume that all of them are well mixed and so if you are going to put your hand in one of the baskets, you will randomly pick out one of these fruits with equal probability. So this basket therefore has six oranges and two apples. This basket here has three apples and one orange. Further assume that your choice of one basket or the other is not equally probable but that picking up, let us say the red basket you pick up with the probability of 0.4, that is 40 percent of the times you will pick the red basket and 60 percent of the times you will pick the blue basket.

So you are not going to pick a basket with equal probability. Notice this notation, probability of B equal to R, this we looked at in the last video. Please notice the random variables that we have here. The random variables we have here are B, which basket we pick and F, which fruit we pick. So instead of the cases which we looked at in the last video, we are actually going to look at a case where you have not just one but two different random variables. That is the basket you pick and further on which fruit you pick within the basket.

(Refer Slide Time: 4:34)

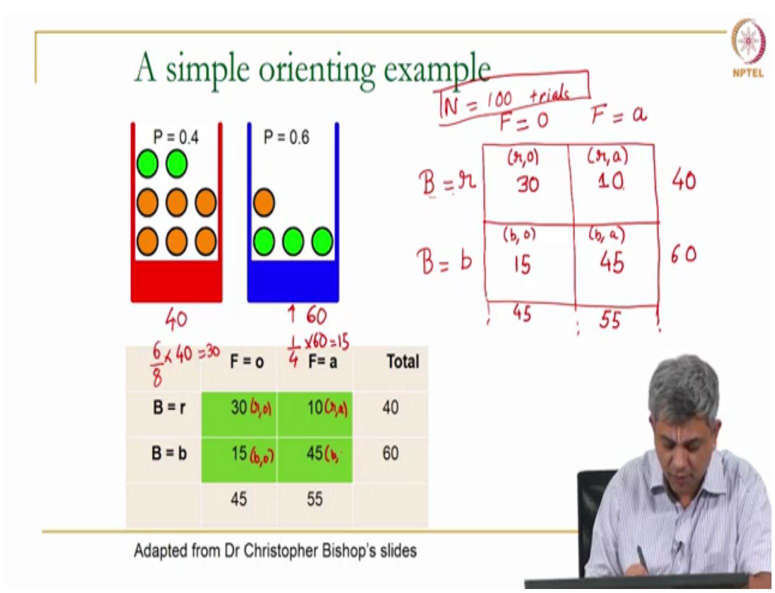


So let us consider this case. Once again the same example. If we look at the random variables in this case, the basket B has the sample space, either you pick blue, sorry, yeah either you pick blue or you pick red. So these are the two possibilities. Amongst fruits you can either pick oranges or you can pick apples. So now we can ask multiple questions. For example, you could ask, what is the probability of picking an orange? Now clearly the probability of picking an orange in this basket is different from the probability of picking an orange from this basket.

But randomly if you just put your hand in one of the baskets and pick a fruit, what is the probability that it is going to be an orange given that I pick red basket with probability 0.4 and blue basket with probability 0.6. This is one simple question we could ask. You could ask slightly more complex question like what is the probability that I picked the red basket given that the fruit I picked was an orange? So this is a classical conditional probability question. You close your eyes, pick up a fruit, it turns out to be an orange.

And now you want to know did you pick it up from the red basket or from the blue basket. You can see that since oranges are more prevalent in the red basket, it might seem like that is a little bit more likely. So we can ask such questions and we can ask far more complex questions. All of these, currently we are only looking at discrete probability examples but all of these are indicative of the kind of questions we will ask later on even within the machine learning context.

(Refer Slide Time: 6:57)



So let us come back here and let us take a case where I take 100 trials. And we are going to assume that the number of cases where you are going to pick red turns out to be exactly 0.4 times 100. This turns out to be true only when actually N tends to infinity but we will assume that everything comes out exactly as if the probabilities are working out as fractions. So let us make a quick table. So if I make 100 trials, so remember I have two rand variables, the basket. So I could have chosen the red basket or I could have chosen the blue basket.

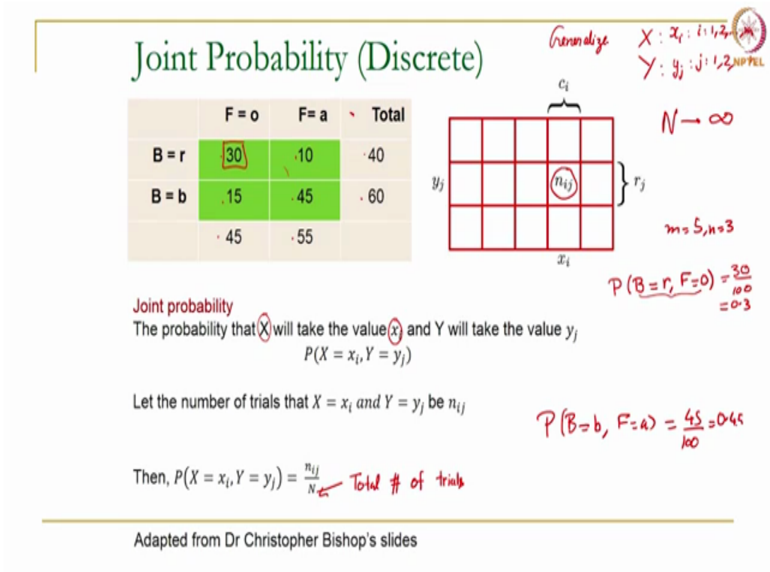
Similarly the fruit could have been an orange or the fruit could have been an apple. So now out of 100 trials, we want to know. So this is the case where the basket I picked was red and the fruit I picked was an orange. This is red and an apple. This is blue, orange, and blue apple. So let us try and find out how many of each of these cases occur. So I know that in the case that I have 100 trials the basket will be red for total of 40 times.

Similarly the basket will be blue a total of 60 times. Now the 40 times that I pick this basket, suppose I want to know in how many of the cases will the fruit be an orange and we can see automatically that assuming it works out exactly according to the probability sixth-eight of the cases which is 30 of the cases you are going to get an orange.

Two-eight of the cases, ten of the cases you get an apple. So a red basket with an apple occurs ten times. Now let us find out a similar case. 60 of the cases are the blue basket, within that an orange comes one-fourth of 60 which turns out to be 15. And we know now that the rest of the 45 cases we must actually be picking an apple. So this is a table which tells you how many times each of these cases occurs. You can also add this up and get the case that out of the 100 times, when I say 100 trials what does it mean? In each trial you pick the basket and chose a fruit.

So amongst those 45 of the times we actually picked an orange and 55 times we actually picked an apple. So we can put this table together in this way, so I will be repeating this table in the future slides. So you can see basket is red, basket is blue and I have written this table out which tells which, how many of these cases, remember each of these actually indicate an intersection of the two cases, basically both these cases occur together.

(Refer Slide Time: 10:38)



So such distributions are called joint distributions. Right now I have written the numbers. The probability obviously of each of these cases is going to be this divided by 100. So you can do 0.3, 0.15, et cetera if you want the probabilities to work out nicely in this case. Now we can generalize this to two variables. Let us say you have a variable X and a variable Y just like in this case we had B and F, basket and fruit.

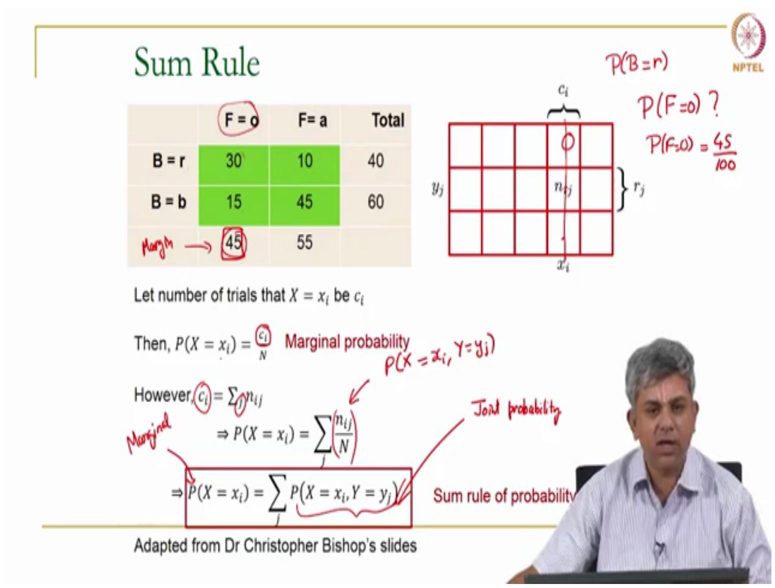
You can have two general variables X and Y. In the example case, the basket B, the random variable B had only two choices, it was either red or it was blue. But you can imagine a case where you have many many more possibilities. So you have Xi and let us say i goes from 1 to some state m. So X instead of two possibilities has many possibilities, let us say m possibilities. And Y also has some n possibilities. In this case we have chosen something like m is equal to 5 and n is equal to 3. But you can obviously choose different numbers.

So your table, instead of a 2 by 2 table which I have shown here, you will have m cross n table of how many times does each case occur. So you take a large number of trials n and really speaking you will get the right fractions only as N tends to infinity unlike the pseudo example I took last time or in this case where I have taken 100 and assumed that it works out. Typically you have to take a very large number n in order for the trials to work out exactly according to their probabilities.

So if we take that, we can define something called the joint probability. Joint probability is the probability that X will take desired value Xi and Y will take some desired value Yj. For example, in our case I could ask something like what is the probability that the basket is red and the fruit is orange. So that would be an example of a joint probability. So you write it under the notation, PX equal to Xi just like here and Y equal to Yj. So if we want that, how would we do it? Let us say I want this. I would say this is 30 divided by the total number of trials which was 100, so this is 0.3.

In the general case let us assume that this box here Xi Yj it has Nij entries. You can think of it as a matrix. Let us say the matrix N has Nij entries there. Then the probability X equal to Xi, Y equal to Yj is Nij by N where N is the total number of trials. So similarly you can ask what is probability, B equal to blue and fruit is an apple, that is going to be 45 by 100. It is 0.45.

(Refer Slide Time: 14:03)



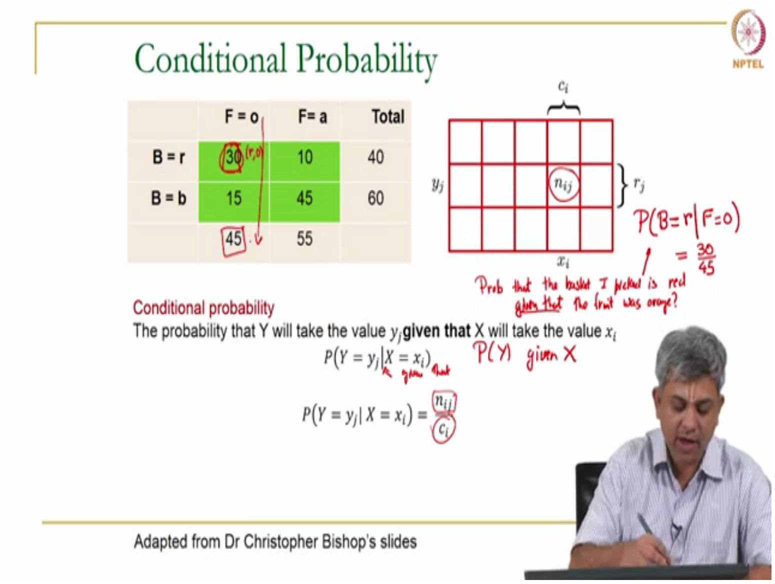
So now using this we come to an important rule called the sum rule. The sum rule asks the question which is, if I do not want a joint probability but I simply want the question, what is the probability that the basket is red? Or I could ask the question, what is the probability that the fruit is an orange? Now you can see this immediately. The probability that the fruit is an orange is going to be 30 of the cases where the fruit was an orange and the basket was red, 15 of the cases where fruit was an orange and the basket was blue, which means a total of 45 cases.

So the probability that the fruit is an orange is going to be 45 divided by 100. So here Ci basically is the sum of this column which is what we got here, regardless of what value of Y was taken. So we take a summation over all the possible values of Y. In this case all the possible baskets that the fruit could have come out of. And the total there is called Ci. And the total number of trials obviously remains the same. So now we automatically see this is called the marginal probability. I will come to the reason for that name shortly.

We can automatically see that Ci basically has to be a summation of Nij over all possible values of j. So j equal to 1, j equal to 3, j equal to 4, so on and so forth till j equal to n, if you sum all of them up, all of the individual boxes up, you are going to get sigma Nij is equal to Ci. So therefore we can write P of X equal to Xi, is equal to sigma of Nij by N. Now this Nij by N you might recall from the last slide is simply each of these is the joint probability. So this is called the sum rule of probability. It is a very important rule. We will be using this multiple times.

This probability here is the marginal probability. This summation is sometimes called the marginalization. So we say the marginal probability, P of X is the marginalization of the joint probability. The terminology might look a little bit confusing but the idea is very very simple, it is just like this. Now why is it called marginal probability? If you notice these numbers, these total numbers are written in the margin. That is the historical origin. Marginalization does not mean anything else, it simply means that the columns or the rows have been summed up and you have put them the total in the margin which is why the total probability is called the marginal probability.

(Refer Slide Time: 17:13)



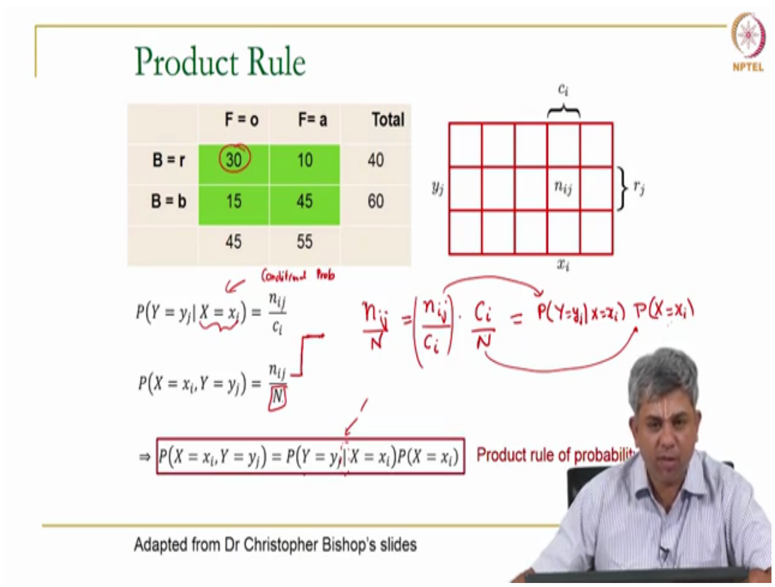
The next idea which you will most probably be familiar with already is the idea of conditional probability. So let us say we ask a question. Instead of just asking what is the probability that the basket is red and the fruit is orange, you can ask a question similar to the one that I asked in the beginning of this video, which is what is the probability that the basket I picked is red given that the fruit was orange? So please understand sometimes people get confused between conditional probability and joint probability. We will see that shortly in the next slide. That, the two are slightly different.

So there is one question which is, what is the probability that the basket I picked was red and the fruit was orange? And the second thing has some information removed, it says that the fruit you picked you finally see is an orange and now you want to know which basket it came out of. Did it come out of the red basket or did it come of the blue basket? We will be using this idea a little bit later just to orient you when you come to interpreting images. So I can ask a question like what is the probability that this is a dog given the certain values of pixels that I am giving. So that is the way we will be utilizing the idea of conditional probability later on.

What is the probability of given output given a certain input, P of Y given X? So this is the way we write it. This is read of probability of Y given X, this is the way we usually say it orally, you should be familiar with this kind of language already. So P of Y is equal to Y given that X is equal to Xi. So in our case this would be written as probability that the basket is equal to red given that the fruit is orange. Now intuitively let us answer this question. So what we said that we are given that the fruit is an orange and out of the trials we know that 45 of those cases are oranges. Now out of those only 30 cases was the basket red which means our probability is going to be 30 over 45.

More generally we would write it as Nij which is equivalent to 30 here, divided by 45 which is Ci which is the column sum. So in general you would write this probability of Y equal to Yj given X equal to Xi as Nij divided by Ci where Ci is the total or the marginal sum of that particular column.

(Refer Slide Time: 20:28)

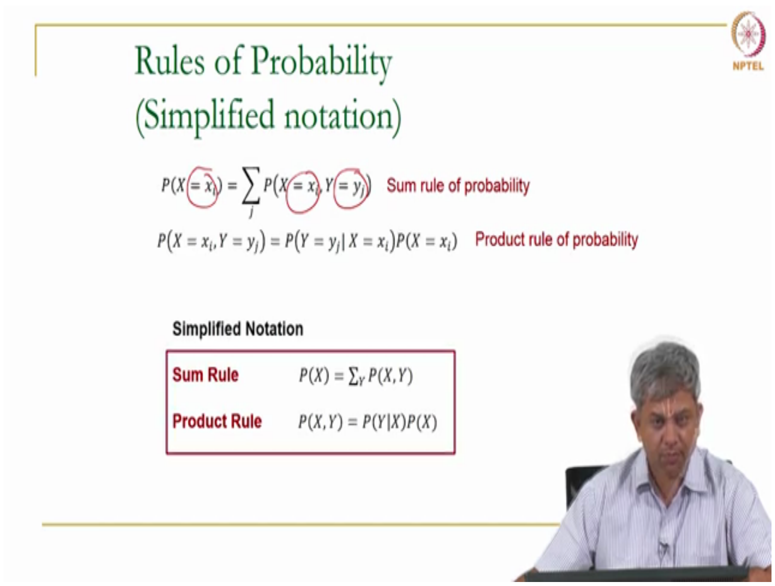


Now here we come to extend this idea to the idea of a product rule. So remember that the conditional probability was Nij by Ci. What was the joint probability? It is Nij by N. You can see that the numerator is the same because we are interested in the same case where the red and the orange occur. But the denominator here when I am talking about joint probability of both these occurring is the total number of trails. Whereas what I know here is a little bit stronger. So I know in this case that the fruit is already an orange.

In this case I do not know what the fruit actually is, so which is why I am dividing by total N. So now if we see this, we can therefore write, let us go from here, Nij by N is equal to Nij by Ci, multiplied by Ci by N. Now this part we have just seen is the conditional probability. This part here, you might recall, is nothing but the marginal probability. Or another way to say it in this case would be probability that the fruit is an orange.

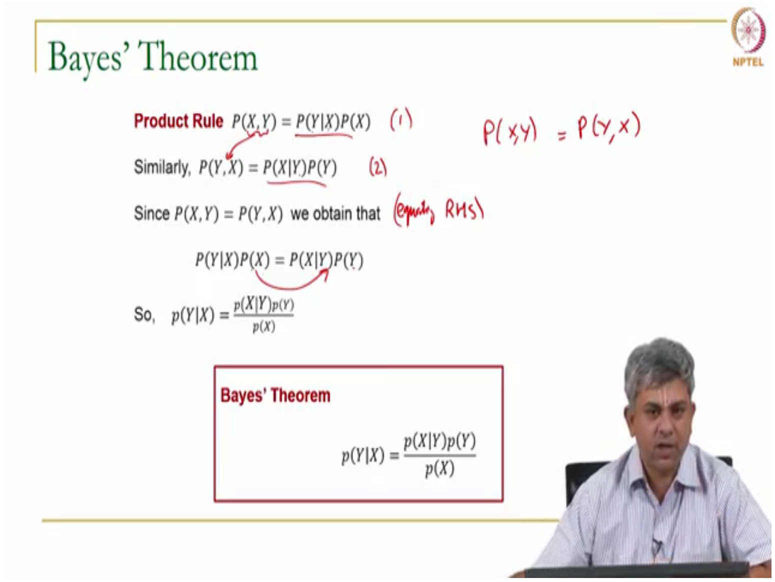
So this rule is called the product rule. Probability that X equal to Xi, Y equal to Yj is probability of Y equal to Yj given X equal to Xi multiplied by this, you can even see people you chosen this notation very carefully. You can even think of this as if it is a division sign to that you can think of this as P of X, Y; P of this by this multiplied by this. So that is just for memory sake, otherwise this rule is called the product rule of probability.

(Refer Slide Time: 22:37)



And these two put together summarize the two important rules of probability which we will be using again and again and again. Practically every theorem in probability can be derived from these two rules, at least the theorems in probability that we will be using. So we simplify the notation. Please notice these Xi and Yj, you do require it in order for rigor but we will not use them, we will actually use a simplified notation. We will simply call this P of X and drop the equal to Xi and equal to Yj. You have to understand what it means according to the context. So this way it is a little bit easier to see. P of X is sigma of P of X, Y and P of X, Y is P of Y given X multiplied by P of X which is the product rule.

(Refer Slide Time: 23:33)



Now we come to Bayes' Theorem which is a simple consequence of the rules that we have seen so far, particularly the product rule. It falls out naturally, you would be familiar with Bayes' Theorem. Even earlier we will be using this several times extensively throughout this course. I will just show a simple derivation right now and we will look at examples in the next video. So let us start with the product rule, P of X, Y is P of Y given X multiplied by P of X. Now you can switch the variables here. So this now becomes P of Y, X. You can now switch X and Y and write this.

Let us say this is equation 1, this is equation 2. But we know that P of X, Y because physically all it means is both the events occur. The basket was red and the fruit was an orange is the same as saying the fruit was an orange and the basket was red. It does not really matter in which order you give joint probability. The variables actually commute. So now if we equate the right hand side, you get P of Y given X, multiplying PX is P of X given Y, multiplying PY and then if we just take this here to the right hand side, you get P of Y given X is P of X, Y, PY divided by PX.

So this is called Bayes' Theorem. It is an extremely useful theorem, sometimes has very counter intuitive results, or at least non-intuitive results if not counter intuitive. We have, we can get non-intuitive results as we will see in the next video. So we will see both examples of the basket problem and another problem the next video just in order to see how Bayes' Theorem can be used.